THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

# QUANTITATIVE METHODS FOR DIFFUSION MEASUREMENTS IN FLUORESCENCE MICROSCOPY

MARCO LONGFILS


Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
AND UNIVERSITY OF GOTHENBURG
Göteborg, Sweden 2019

This work has been financially supported by the Swedish Foundation for Strategic Research (SSF).

**Quantitative methods for diffusion measurements in fluorescence microscopy**
MARCO LONGFILS

ISBN 978-91-7905-112-9



Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 4579
ISSN 0346-718X

Department of Mathematical Sciences
Chalmers University of Technology
and University of Gothenburg
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 35 28
Author e-mail: longfils@chalmers.se

**Cover:** Visualization of a sequence of ten consecutive frames of a single particle tracking simulation of diffusing particles. The method used in Paper I and Paper II belongs to the family of single particle tracking techniques. The ten frames are overlapping in this figure and the color, from blue to red, represents the time.

Typeset with LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2019

# Quantitative methods for diffusion measurements in fluorescence microscopy

Marco Longfils

*Department of Mathematical Sciences*
*Chalmers University of Technology and University of Gothenburg*

## Abstract

In this work, statistical methods are developed for mapping mass transport locally based on images collected using a confocal laser scanning microscope. Besides presenting raster image correlation spectroscopy as an established method in fluorescence microscopy, we introduce a single particle tracking method which takes advantage of the raster scanning of the image in a confocal microscope. In single particle tracking, particles are identified and followed in consecutive frames of a video to measure their diffusive mobility. Both a maximum likelihood and a centroid-based method have been developed to locate the particles and hence to estimate the diffusion coefficient. The method is generalized to analyse mixtures of particles having different diffusion coefficients. The proposed method allows us to study the entire distribution of diffusion coefficients, enabling the characterization of heterogeneous systems. Motivated by experiments with particle mixtures, we investigate the use of cross-validation to perform model selection, i.e. to select the number of mixture components, and compare it to some existing model selection criteria. In the specific case of normal mixtures, we prove a bound on the error between the cross-validated conditional risk and an oracle benchmark conditional risk, which assumes the knowledge of the true density generating the data. Furthermore, a detailed statistical analysis of the raster image correlation spectroscopy method is presented, uncovering the relationship between molecular and experimental parameters and the estimated diffusion coefficient. We propose a statistical method to compare different experimental conditions and apply it to find the optimal parameters to perform an experiment.
The methods and models investigated and developed in this thesis are of general interest. In particular, the quantitative methods considered to study confocal images can be used in a wide range of applications, while the use of cross-validation to perform model selection of mixture models is a valuable contribution to the statistical literature.

**Keywords:** Confocal laser scanning microscopy, diffusion, correlation spectroscopy, raster scan, single particle tracking, mixture models, cross-validation.

# List of appended papers

The following papers are included in this thesis:

Paper I. **Longfils, M.**, Schuster, E., Lorén, N., Särkkä, A., and Rudemo, M. (2017). Single particle raster image analysis of diffusion. *Journal of Microscopy*, 266, 3-14.

Paper II. **Longfils, M.**, Röding, M., Altskär A.; Schuster, E., Lorén, N., Särkkä, A., ands Rudemo, M. (2018). Single particle raster image analysis of diffusion for particle mixtures. *Journal of Microscopy*, 269, 269-281.

Paper III. **Longfils, M.**, Röding, M., Lorén, N., Särkkä, A., and Rudemo, M. Identification of mixture models with an application to diffusing particles. *Manuscript*.

Paper IV. **Longfils, M.**, Smisdom, N., Ameloot, M., Rudemo, M., Lemmens, V., Fernández, G.S., Röding, M., E., Lorén, N., Hendrix, J. and Särkkä, A. Raster image correlation spectroscopy performance evaluation. *Submitted*.

My contribution to the appended papers:

Paper I: I co-developed and implemented the single particle raster image analysis. Moreover, I carried out the simulation study as well as the analysis of the experimental data. I also participated in the data collection and did most of the writing for the publication.

Paper II: I implemented both single particle raster image analysis and raster image correlation spectroscopy in the case of mixtures of particles. Moreover, I carried out the simulation study as well as the analysis of the experimental data. I also did most of the writing for the publication.

Paper III: I proved the finite sample results for cross validation. Moreover, I carried out the simulation study as well as the analysis of the experimental data. I also did most of the writing for the publication.

Paper IV: I developed the theory and the calculations necessary to perform the method Raster image correlation spectroscopy performance evaluation. I implemented the programs in Matlab as well as built the graphical user interface. I carried out the simulations and helped with the analysis of the experiments. I did most of the writing of the publication.

**Publications not included in this thesis:**

Eliasdottir, O., Hildeman, A., **Longfils, M.**, Nerman, O., and Lycke, J. (2018). A nationwide survey of the influence of month of birth on the risk of developing multiple sclerosis in Sweden and Iceland. Journal of Neurology, 265, 108-114.

Andersen, O., Hildeman, A., **Longfils, M.**, Tedeholm, H., Skoog, B., Tian, W., Zhong, J., Ekholm, S., Novakova, L., Runmarker, B., Nerman, O., and Maier, S.E. (2018). Diffusion tensor imaging in multiple sclerosis at different final outcomes. Acta Neurologica Scandinavica, 137, 165-173.

# Acknowledgements

First and foremost, I would like to thank my supervisor Aila Särkkä for her continuous support in every aspect of my studies. I dropped by your office far too many times and you always had time for me. I would also like to thank Mats Rudemo, for the very interesting discussion we had about my project. Aila and Mats, what you did for me is way more than being supervisors. Next, I would like to thank Magnus Röding for encouraging me to become a better researcher. Magnus, your invaluable comments turned my sloppy manuscripts into articles. Aila, Magnus, and Mats, my ideas were just pieces and thanks to you I could complete the puzzle. Furthermore, I thank Sergei Zuyev, for teaching me most of what I know about probability and the exciting discussions we had.

I would also like to thank Magnus, Sandra, Cecilia, Henrike, and Torben for the nice fikas we had within the SSF project. A big thanks goes to Niklas Lorén, Erich Schuster, and Annika Altskär from RISE for creating a fascinating and challenging interdisciplinary research environment, and for providing interesting microscopy data without which my work would not be possible. I am also grateful to my colleagues for making the department an enjoyable workplace. I am grateful to Marcel Ameloot, Nick Smisdom, Jelle Hendrix, Veerle Lemmens, Hannelore Bové, and Eli Slenders for the time I spent in Hasselt. I have grown so much and I have matured, thanks to you, experimentally speaking in only two months, and you made me feel home. I am indebted to Petter Mostad, who had infinite patience and coached me to become a better teacher. Olle Nerman, I will miss the times you knocked on my door for the consultancy projects. I tried to learn as much as possible from one of the most knowledgeable statisticians.

Let me pay homage to an important person who supported me constantly. I will put it in ink: you are a goat, Eminem.

Last, but not least, I am grateful to my parents and my brother for their love and constant support. Everything I have reached so far I owe to you.

Financial support from the Swedish Foundation for Strategic Research, SSF, is highly appreciated.

*This thesis is dedicated to my parents and my brother.*
*Whenever you need me, I'm never too far.*

I bully myself 'cause I make me
do what I put my mind to.

---

*EMINEM*
*Rap God, 2013*

# Contents

# 1 Introduction

In many applications, ranging from packaging materials to pharmaceuticals, to be able to design biomaterials with tuned mass transport functionalities is essential. Therefore, understanding the microstructure - mass transport relationship is highly important. In order to successfully create such materials, measurement methods need to resolve the mass transport (diffusion in our case) properties at the length scales of the material structures. This requires performing measurements with (sub-)micrometer spatial resolution. The aim of this work is to develop a new, high-accuracy statistical method to map mass transport heterogeneity at a (sub-)micrometer scale and further promote existing microscopy methods to determine mass transport.

In this work, we concentrate on pure diffusion and how to estimate diffusion coefficients, both when there is only one (monodisperse) and when there are several (polydisperse) diffusion coefficients. The organization of this thesis is as follows: Section 2 gives an introduction to diffusion, which is the main interest of the appended papers. The principles of confocal microscopy, which is used to acquire the data, are briefly introduced in Section 3. Section 4 provides an overview of two large families of available methodologies employed to study diffusion, image correlation spectroscopy and single particle tracking. In particular, as the method called single particle raster image analysis is the main focus of two of the appended papers, only a brief description of this method will be given in the introduction of this thesis. Some more details are provided for the raster image correlation spectroscopy technique. In general, when studying a polydisperse system, the number of different particle types is unknown and we choose between models with different numbers of components. Section 5 contains an introduction to the model selection problem. In Section 6, the appended papers are summarised and in Section 7, possible topics of study in future work are presented.

# 2 Diffusion

Diffusion is the migration or movement of particles due to random motion driven by thermal energy. There are three main perspectives on how we can look at diffusion: Fick's law, the Wiener process, and the Einstein-Smoluchowski relation. To describe the three points of view, we start by considering pure diffusion of particles with a single diffusion coefficient in a homogeneous medium. In this case, the diffusion coefficient of all the particles will be the same, as the system is monodisperse, and constant over space, as the medium in which diffusion takes place is homogeneous. Let $C(r,t)$ and $\delta C(r,t)$, respectively, be the concentration of particles and the deviation from the average concentration at the position $r \in \mathbb{R}^3$ and time $t \in [0,\infty)$. The temporal evolution of such a system is described by Fick's (second) law of diffusion,

$$\frac{\partial u(r,t)}{\partial t} = D\nabla^2 u(r,t), \tag{2.1}$$

where $D > 0$ is the diffusion coefficient and $\nabla^2$ is the Laplacian operator. Fick's law is satisfied for both the concentration $C(r,t)$ and the deviation from the average concentration $\Delta C(r,t)$ by linearity, and predicts how these quantities change with time. The physical properties of diffusion are characterised by a density function $P(r,t)$, called the propagator. The propagator specifies the probability density of finding a particle located at $r$ at time $t$ when the particle was at the origin at time zero. The propagator is given by

$$P(r,t) = \frac{1}{(4\pi Dt)^{\frac{3}{2}}} e^{-\frac{\|r\|^2}{4Dt}}. \tag{2.2}$$

The propagator fully describes the type of movement exhibited by the particles and is directly involved in the correlation function used by raster image correlation spectroscopy. For flow with a velocity vector $V = (V_x, V_y, V_z)$, the propagator takes the form

$$P(r,t) = \delta(r_x - V_x t)\delta(r_y - V_y t)\delta(r_z - V_z t), \tag{2.3}$$

where $\delta(x-y)$ equals one if $x = y$ and zero otherwise. There are more propagators which have been used for different modes of motion, e.g. for directed diffusion and anomalous diffusion. The former is the superposition of flow and pure diffusion, while the latter generally defines a deviation from pure diffusion. Anomalous diffusion is characterised by the displacement having a second

moment which follows a power law $\sim t^\alpha$ as a function of time , and is usually classified as subdiffusion for $\alpha < 1$ or superdiffusion for $\alpha > 1$. This type of motion has been observed in cell membranes as a result of both obstacles and binding kinetics. As described in Bouchard & Georges (1990), one way to model anomalous diffusion is to consider particles performing a random walk where the jumps are drawn from a broad distribution or show long range correlation. Thus, the usual central limit theorem does not hold anymore and the law of Brownian motion, corresponding to pure diffusion, is not valid.

Fick's law describes the macroscopic properties of diffusion, as it defines how the concentration of particles changes in time. On the other hand, the following interpretation of diffusion in terms of the Wiener process provides a microscopic view of the process, as it gives a description of diffusion in terms of the motion of the single particles. Molecules undergoing diffusion are mathematically modelled as particles moving according to a Brownian motion, c.f. Equation (2.2), where the variance of the Gaussian increments is proportional to the interval of time considered. The proportionality constant is the diffusion coefficient (up to a dimensionality constant). Formally, consider $n$ diffusing particles and let $X_i(t) = (X_i^1(t), ..., X_i^d(t)),\ i = 1, ..., n$, denote the vector of the position in $\mathbb{R}^d$ of the $i$-th particle at time $t$. Then, $X_i^1, ..., X_i^d$ are independent translated copies of Wiener processes $W$ defined by:

1. $W_i(0) = 0$;
2. $W_i(t) - W_i(s) \sim N(0, 2D_i(t-s))\ \forall t > s \geq 0$, where $D_i$ is the diffusion coefficient of the $i$-th particle;
3. Increments of $W_i$ for nonoverlapping time intervals are independent.

The last perspective on diffusion is given by the Einstein-Smoluchowski relation. It was first derived by Einstein (1905) and a year later independently by Smoluchowski (1906), and it links the macroscopic diffusion coefficient D to the microscopic information about the mean square displacement,

$$\mathbb{E}\left[\|X(t+\Delta t) - X(t)\|^2\right] = 2dD\Delta t. \tag{2.4}$$

The above representations of diffusion are exploited by image correlation techniques and single particle methods. In the first family of techniques, as particles appear as bright spots in the image due to the fluorescent labelling, the correlation between and/or within images is coupled to the probability of finding the same particle again at some spatiotemporal lag, which in turn is related to the propagator. In single particle methods, the displacements $X_i(t+\Delta t) - X_i(t)$ are directly estimated for some fixed temporal lag $\Delta t$, for example the time between consecutive images. Then, the diffusion coefficient can be estimated from the second moment of the displacements.

In a more general case, particles can interact with each other or spatially with particular structures like binding sites. As an example, let us consider a solution of particles having $m$ distinct diffusion coefficients and denote by $C_j(r,t)$ and $\Delta C_j(r,t)$, respectively, the concentration of the $j$-th component (particle type) and the deviation from the average concentration, for components $j = 1, \ldots, m$. Moreover, denote by $D_j$, $j = 1, \ldots, m$, the diffusion coefficient of the $j$-th component. Near equilibrium, the system evolves according to the so-called reaction-diffusion equation

$$\frac{\partial \Delta C_j(r,t)}{\partial t} = D_j \nabla^2 \Delta C_j(r,t) + \sum_{k=1}^{m} K_{jk} \Delta C_k(r,t), \tag{2.5}$$

where the first term on the right hand side accounts for diffusion and the second describes changes due to interaction, and where $K_{jk}$ are the chemical rate constants. In this work, we will restrict ourselves to diffusion, leaving interaction for future studies.

# 3 Microscopy data

All data considered here were collected with a confocal laser scanning microscope (CLSM), see Pawley (2006) for a comprehensive introduction to the subject. In CLSM, a laser beam is passed through an illumination aperture which is then focused by an objective lens into a small area of the sample, see Figure 3.1. If fluorophores are present and illuminated with the proper wavelength, they emit light. This light then passes through a semi-transparent mirror, the dichroic mirror, towards the detection system. At this point, light passes through the emission filter, which separates the fluorescent light from the laser light reflected by the sample. For every pixel in the region of interest, the number of photons emitted from the fluorophores in the sample or an intensity value is recorded by a photon counter or a photo multiplier tube. A pinhole aperture in front of the detector is used to exclude fluorescence from the out-of-focus planes.

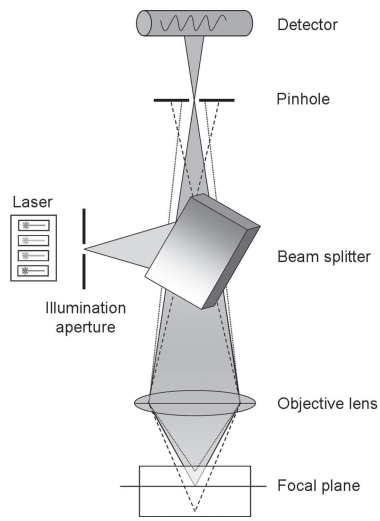


Figure 3.1: Schematic illustration of confocal laser scanning microscopy.

In Figure 3.1, the light beams from the sample that come from the out-of-focus planes, represented as dashed and dotted lines, are stopped by the pinhole and

not collected by the detector. Hence, confocal microscopy provides a "well-isolated" plane. Confocal microscopes allow detection of fluorescent molecules with good spatial resolution. In the experiments of papers I-III, fluorescent microspheres have been used as probes to study diffusion. We considered microspheres with four different diameters (100, 175, 500, 1000 nm) where the smallest size covers the subresolution domain, while the largest size is nearly the size of a living cell. Within each size considered, the standard deviation of the diameter is remarkably small, typically around 2-3% of the size, which allows us to consider the particles as identical in terms of size and shape. The homogeneity of the microspheres is fundamental to ensure that any variation in their motion is due to the surrounding structure. Moreover, the beads have been stained with four different fluorescent dyes. Thus they will be visible only if excited with one of the corresponding four well-separated wavelengths. In applications, we can use different colors to label the structure or important immobile features in the sample and the particles. By observing their motion using different detectors, we can easily separate the background from the diffusing microspheres. The fluorescent dye is used to stain the particles in such a way that the fluorophore distribution is uniform over the volume of each microsphere. In a confocal image, an immobilized fluorescent microsphere appears as a bright round object, the radius of which depends on the distance of the particle to the focal plane and the size of the particle. The closer the particle is to the focal plane, the larger the radius will be, see for example the top left plot in Figure 4.3. In paper IV, solutions with different fluorescent dyes, Atto488-COOH or enhanced green fluorescent protein dissolved in a PBS or sucrose buffer, have been used.

## 3.1 Photon detection process

The process leading to the pixel intensity in the confocal microscope is termed photon detection process. The photon detection process can be modelled as a Cox process (Cox, 1955). To explain this concept further, we need to introduce some basic definitions. We present here a brief introduction to the theory of point processes, and refer to (Diggle, 2013) for a comprehensive presentation of the subject. Let $N$ be the family of all subsets in $\mathbb{R}^d$ that satisfy the two following conditions:

1. an element $\psi \in N$ is locally finite, i.e. each bounded subset of $\mathbb{R}^d$ can only contain a finite number of elements of $\psi$;

2. $\psi$ is simple, so if we denote $\psi = \{x_i,\ i = 1, 2, ...\}$ then $x_i \neq x_j$ if $i \neq j$.

A point process $\Phi$ on $\mathbb{R}^d$ is a random variable taking values in the measurable space $[N, \mathcal{N}]$, where $\mathcal{N}$ is the smallest $\sigma-$algebra that makes all mappings $\phi \rightarrow$

$\phi \cap B$ measurable for all bounded Borel sets $B$. More intuitively, it is a random choice of one of the elements in $N$. To characterize a point process we often use its intensity measure $\Lambda$, which describes the expected number of points in a Borel set $B$

$$\Lambda(B) = \mathbb{E}[|\Phi \cap B|],$$

where $|A|$ denotes the number of elements in a set $A$. We restrict ourselves to the case where $\Lambda$ admits a density $\lambda$ with respect to the Lebesque measure on $\mathbb{R}^d$, called the intensity function. Then, $\lambda(x)dV$ is interpreted as the infinitesimal probability that there is a point of $\Phi$ in a region of infinitesimal volume $dV$ located at $x$. The simplest point process is the Poisson point process (PPP). A PPP $\Phi$ with intensity measure $\Lambda$ is characterized by two properties. First, the number of points of $\Phi$ in any bounded Borel set $B$ follows a Poisson distribution with mean $\Lambda(B)$. Second, the random variables that count the number of points of the process in $k$ disjoint Borel sets are independent. Consider a subset $M \subset \mathbb{R}^d$ and a locally integrable, non-negative random field $Z = \{Z(u) : u \in M\}$. A point process $F$ is said to be a Cox process driven by the random intensity function $Z$ if, conditionally on $Z = z$, $F$ is a PPP with intensity function $z$. Let $\Phi(t) = \{X_1(t), X_2(t), \ldots\}$ be the PPP which models the random positions $X_1(t), X_2(t), \ldots$ of the fluorescent particles in $\mathbb{R}^3$ at time $t$ and define the following random intensity function

$$Z(u) = \lambda T \sum_{X \in \Phi} I(u - X). \tag{3.1}$$

Here, $\lambda$ is the photon yield of the particle, $T$ is the integration time (pixel dwell time), and $I$ is the excitation light intensity profile of the laser, for a location $u = (u_x, u_y, u_z)$, which is given by

$$I(u) = I_0 \exp\left\{ -\frac{2(u_x^2 + u_y^2)}{w^2} - \frac{2u_z^2}{\alpha^2 w^2} \right\}. \tag{3.2}$$

The exponential term in the right hand side of Equation 3.2 is referred to as the (scaled) point spread function (PSF), $w$ is the lateral waist or radius of the PSF, and $I_0$ accounts for the laser power.

The intensity $F(u,t)$ for the pixel at position $u$ in the image at time $t$ represents the number of photon with arrival time between $t$ and $t+T$ detected by the confocal microscope. $F(u,t)$ is, conditionally on the realization $Z = z$, a Poisson random variable with parameter $z(u)$. Hence, the fluorescence intensity in a pixel can be described by a Cox process driven by $Z$ as expressed in Equation (3.1). In Figure 3.2, a realization of the point process $\phi$, the corresponding intensity function $z$, and the confocal image $F$ are shown when the particles have fixed locations.

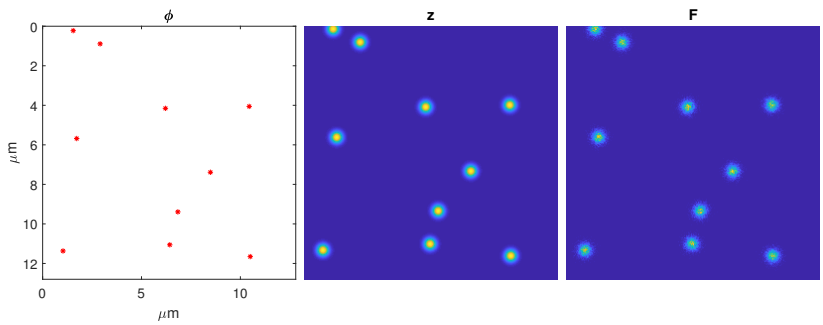


Figure 3.2: Left: A realization of the point process $\phi$ describing the positions of the fluorescent molecules; Middle: Realization $z$ of the random intensity function defined in Eqaution 3.1 for the realization of $\phi$ showed on the left figure; Right: Image obtained by the confocal microscope.

In general, the particles will diffuse over time. Consider two times $t_1 \leq t_2$ and denote by $\Phi(t_1)$, and $\Phi(t_2)$ the particle positions at the two time points. Then, $\Phi(t_2)$ is a random displacement of each of the points in $\Phi(t_1)$ according to a normal distribution with variance proportional to the diffusion coefficient of the particles. Consequently, the corresponding intensities $Z(t_1)$ and $Z(t_2)$, and the photon counts $F(u, t_1)$ and $F(u, t_2)$ will change. The fluctuations in the pixel intensity $F(u, t)$ due to the movement of the particles will be the basis to studying diffusion. In Figure 3.3, two successive confocal images are shown, where the particles diffuse between frames, but can be considered immobile within each frame.

The presented modelling paradigm was first exploited in (Koppel, 1974), and later in (Qian, 1990) to study asymptotic properties of the statistical accuracy of Fluorescence Correlation Spectroscopy. A more extensive and rigorous treatment of the photon detection process can be found in (Saleh, 1978).

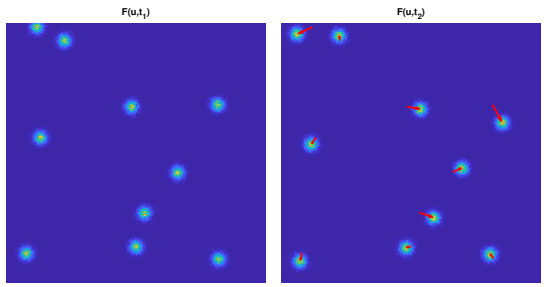


Figure 3.3: Two successive confocal images of fluorescent particles, with the displacements indicated by red arrows.

# 4 Methods to estimate diffusion from microscopy data

In this work, we focus on two major methodologies to study diffusion, the Image Correlation Spectroscopy (ICS) and Single Particle Tracking (SPT) techniques. Another noteworthy method is Fluorescence Recovery After Photobleaching (FRAP) (Lorén *et al.*, 2015) which was first used to analyze the mobility of individual molecules within a cell membrane. In FRAP, a fluorescent probe is introduced in the sample, a cell or a soft biomaterial. Then, a high intensity laser bleaches rapidly the fluorescence in the region of interest and a sequence of images is collected to follow the recovery rate of the fluorescence. Over time, non-bleached probes will diffuse into the region of interest, while the bleached ones will diffuse out of it. Thus, information about diffusion can be retrieved from the recovery of the fluorescence.

The predecessor of ICS methods, namely Fluorescence Correlation Spectroscopy (FCS) is also worth mentioning. In an FCS experiment, a small volume of the sample is illuminated by a stationary light source and the fluorescence from particles is recorded. Since particles are allowed to diffuse in and out of the observed volume and may undergo chemical and physical processes, fluctuations in the signal will arise. By recording the fluorescence intensity over a time period, a time series will be generated. The temporal autocorrelation function of this time series will be distinct for different types of motion of the particles and interactions like binding. Thus, by analyzing the shape of the autocorrelation function we can determine the behaviour of the particles in the sample and estimate parameters of interest like the diffusion coefficient or the average binding time.

## 4.1 Image Correlation Spectroscopy

We present here a brief overview of Image Correlation Spectoscopy and introduce Raster Image Correlation Spectroscopy (RICS) in more detail. ICS is a unifying term for a group of fluorescence fluctuation spectroscopy techniques based on the analysis of fluorescence microscopy image data. ICS methods are subdivided according to whether fluorescence fluctuation information in space and/or time is analysed within the image series. Temporal ICS (TICS) (Kulkarni *et al.*, 2005) analyses fluorescence fluctuations in time recorded in the pixels of an image time series. Spatiotemporal ICS (STICS) (Hebert *et al.*,

2005) considers information in both space and time. An innovative method is Raster scan ICS (RICS) (Digman *et al.*, 2005), (Brown *et al.*, 2008), (Gielen *et al.*, 2009), which like STICS considers spatiotemporal correlations, but gains access to a faster timescale by exploiting the rapid pixel-to-pixel sampling in a laser scanning microscope. We should point out that many other methods fall under the ICS family, as kICS (k-reciprocal Image Correlation Spectroscopy) (Kolin *et al.*, 2006), ICCS (Image Cross-Correlation Spectroscopy) (Comeau *et al.*, 2006) and variants of them. All variants of ICS are based on an image or image time series recorded using fluorescence microscopy, such as confocal laser scanning microscopy (CLSM) or Total Internal Reflection Fluorescence microscopy (TIRF). In all pixels of an image, the output of the photomultiplier tube or bin counts from a charge-coupled device camera are registered. For example, in the case of a photon counting detector, the pixel intensity represents an actual count of detected photons. The key feature that all ICS methods take advantage of is that the intensity of a point fluorescent source will be spread out upon detection due to the diffraction of light. The diffraction pattern is described by the point spread function (PSF). The PSF is assumed to be a three-dimensional Gaussian function for a confocal microscope with different axial (z-direction) and lateral (xy-plane) standard deviations, see Equation 3.2. Thus, spatial correlation will be introduced between adjacent pixels of the image. The effect of the PSF is shown in Figure 4.1.

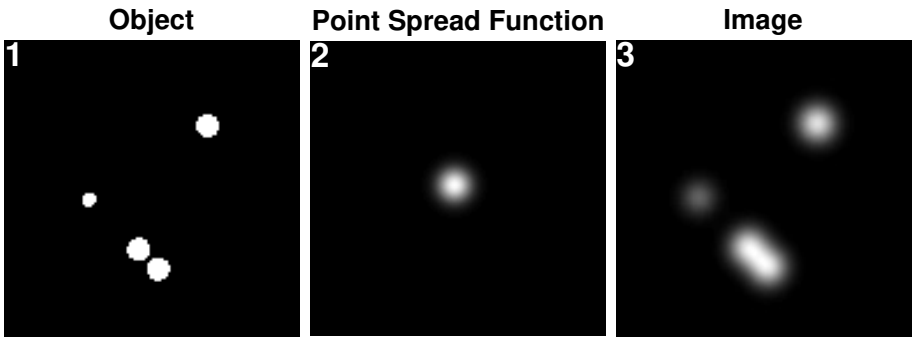


Figure 4.1: Image describing the effect of the point spread function. 1: the object of interest; 2: the point spread function. 3: the image as recorded by the microscope, where the image is the result of the convolution of the other two images.

### 4.1.1 Raster Image Correlation Spectroscopy

In this section, we describe RICS in more detail, and in particular, we focus on the case of line scanning, while many considerations apply also to the case of circular scanning. In RICS, each image is scanned pixel-by-pixel and line-by-line through the movement of the focal observation volume according to a raster pattern. This particular sampling pattern introduces time information within the image. Scanning of the sample is executed as shown in Figure 4.2. The observation volume is placed on the first (from left to right) pixel of the image which is scanned. Then, after the pixel dwell time $\tau_p$, the second pixel in the first line is scanned. Scanning pixel-by-pixel, the first line of the image will be collected. In the next step, after the line time $\tau_l$, the observation volume is retraced to the beginning of the second line of pixels. At this point, the second line is recorded, and by iterating this process the whole image is sampled. Typically, in a RICS measurement, adjacent pixels in the $x$-direction are scanned within a microsecond, and adjacent pixels in the $y$-direction are scanned within a millisecond.

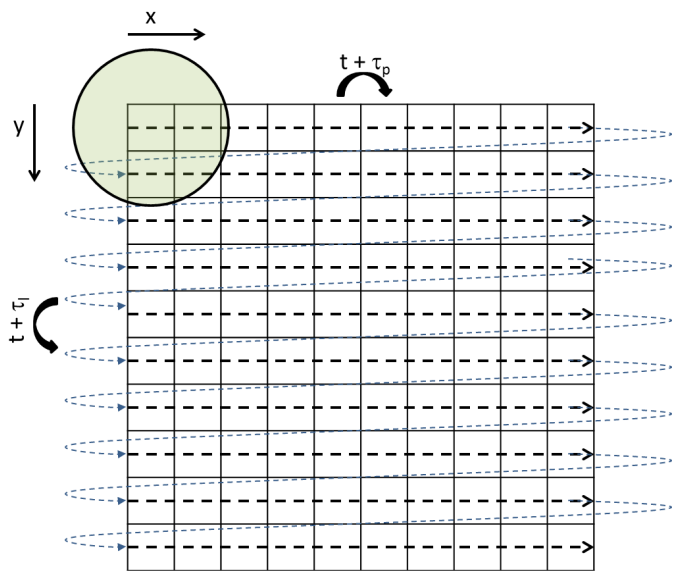


Figure 4.2: Movement of the scanning beam according to the raster scan pattern used in RICS. The scanning time between adjacent pixels in the $x$- and $y$-directions are $\tau_p$ and $\tau_l$, and $\tau_p \ll \tau_l$.

In all correlation spectroscopy techniques, with simple modifications whenever only temporal or only spatial information is analysed, the signal fluctuation with respect to the average is calculated as

$$\Delta F(r,t) = F(r,t) - \langle F(r,t)\rangle,$$

where $F(r,t)$ is the signal in $r$ at time $t$, $\Delta F(r,t)$ is the fluctuation of the signal and $\langle\cdot\rangle$ denotes averaging. In the case of photon counting detectors, the signal $F(r,t)$ recorded by the microscope represents the count of detected photons, which justifies the use of the same notation as in Section 3.1. The normalised correlation of the fluctuations, $G(\rho,\tau)$, is given by:

$$G(\rho,\tau) = \frac{\langle \Delta F(r,t)\Delta F(r+\rho,t+\tau)\rangle}{\langle F(r,t)\rangle^2} = \frac{\langle F(r,t)F(r+\rho,t+\tau)\rangle}{\langle F(r,t)\rangle^2} - 1$$

where $\rho = (\rho_x, \rho_y)$ and $\tau$ are the spatial and temporal shifts. It should be noted that $G(\rho,\tau)$ is not exactly a correlation function, but a normalized covariance function where the maximum of $G(\rho,\tau)$ scales as the inverse of the average number of particles $\langle N\rangle$ in the observation volume. However, in the literature it is referred to as a correlation function, so we will use this name. The correlation function for RICS in the case of pure diffusion, where the lags are $\tau = \tau_p|\xi| + \tau_l|\psi|$ and $\rho_x = S\xi$, $\rho_y = S\psi$, where $S$ is the pixel size and $\xi$ and $\psi$ are the $x$- and $y$-axis spatial increments in number of pixels, is given by

$$\begin{aligned} G(\xi,\psi) =& \frac{1}{\langle N\rangle} e^{\left[-\frac{(S\xi)^2+(S\psi)^2}{w_0^2+4D|\tau_p\xi+\tau_l\psi|}\right]} \left(1+\frac{4D|\tau_p\xi+\tau_l\psi|}{w_0^2}\right)^{-1} \\ & \times \left(1+\frac{4D|\tau_p\xi+\tau_l\psi|}{w_z^2}\right)^{-\frac{1}{2}}. \end{aligned} \tag{4.1}$$

Some examples of such correlation functions are plotted in Figure 4.3 and more details are provided in the papers.

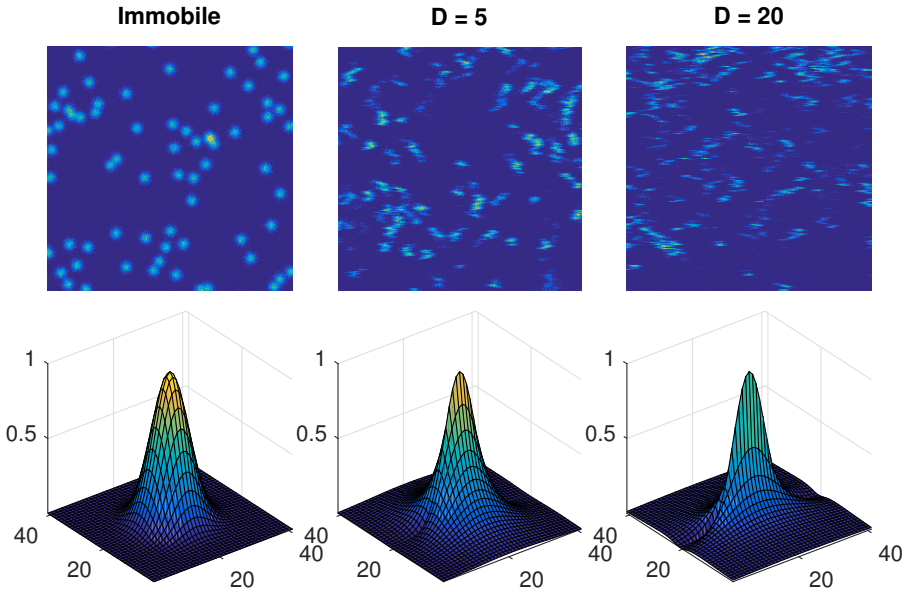


Figure 4.3: Top: Examples of typical RICS images for immobile particles (left), diffusing particles with $D = 5\,\mu m^2\,s^{-1}$ (middle), and diffusing particles with $D = 20\,\mu m^2\,s^{-1}$ (right). Bottom: corresponding theoretical correlation functions for the different cases.

Assume we have $n$ images with resolution $K \times K$ from which we want to estimate diffusion. Let $C(\xi, \psi, j)$ be the empirical correlation function relative to a shift of $\xi$ pixels in the $x$-direction and $\psi$ pixels in the $y$-direction, $1 \leq \xi, \psi \leq K$, of the $j$-th image, $1 \leq j \leq n$. In RICS, the estimation procedure follows the following steps:

1. Compute $C(\cdot, \cdot, j)$ for all $1 \leq j \leq n$ via the Fast Fourier Transform algorithm;

2. To reduce the effect of noise, compute the average empirical correlation function of the stack of images

$$\hat{C}(\xi, \psi) = \frac{1}{n} \sum_{j=1}^{n} C(\xi, \psi, j)$$

3. Consider the following theoretical correlation function depending on the vector of parameters $\theta = (\langle N \rangle, D, O)$, respectively the average number of particles in the observation area, the diffusion coefficient and the offset of the correlation function:

$$G(\xi,\psi,\theta) = \frac{1}{\langle N \rangle} e^{\left[-\frac{(S\xi)^2+(S\psi)^2}{w_0^2+4D|\tau_p\xi+\tau_l\psi|}\right]} \left(1+\frac{4D|\tau_p\xi+\tau_l\psi|}{w_0^2}\right)^{-1} \times \left(1+\frac{4D|\tau_p\xi+\tau_l\psi|}{w_z^2}\right)^{-\frac{1}{2}} + O \tag{4.2}$$

where $\tau_p, \tau_l$, and $S$ are, respectively, the pixel dwell time, line time and pixel size.

4. Define the estimate $\hat{\theta}$ as the weighted least squares estimate of $\theta$, i.e.

$$\hat{\theta} = \arg\min_{\theta} \sum_{\xi,\psi} w(\xi,\psi) \left[G(\xi,\psi,\theta) - \hat{C}(\xi,\psi)\right]^2,$$

where the weights $w(\xi,\psi) = \left(\mathrm{Var}(\hat{C}(\xi,\psi))\right)^{-1}$ are computed from the set of independent images.

## 4.2 Single Particle Tracking

Single Particle Tracking (SPT) was first introduced by Perrin (1909). Since then, many variants of this method have been introduced. However, they share the goal of investigating mass transport and the same measure of mass transport properties, even though the estimation techniques are different. One of the main advantages of SPT is that it gives access to the entire distribution of diffusion coefficients and subpopulations of particles, while other methodologies like FRAP or ICS average the behaviour of hundreds or thousands of diffusing particles. In SPT, a video or a sequence of frames is employed to track the motion of single particles. Here, a "particle" may be anything from a single molecule to a macromolecular complex or microsphere. Typical particles used are fluorescent particles, such as latex beads or gold nanoparticles. The two main steps of the image analysis for SPT are: (i) particle detection, in which bright spots that stand out from the background are identified in some way and their positions estimated in every frame of the video, and (ii) particle linking, in which the detected spots are connected from one frame to the next to form tracks. Some examples of algorithms to localize particles are the centroid algorithm, where the center of mass of the particle is used as a computationally simple and efficient estimate of its position, and the Gaussian fit algorithm,

where a 2D or 3D Gaussian curve is fitted to the profile of the particle, and the mean provides a measure of the position. From the estimated trajectories one can extract the mean square displacement (MSD) which contains information about the type of motion. Let $x(t) \in \mathbb{R}^d$ be the position of the particle at time $t$. The MSD is defined as follows:

$$MSD(t) = \mathbb{E}\left[\|(x(s+t) - x(s))\|^2\right], \tag{4.3}$$

where $\|\cdot\|$ denotes the Eucledian norm in $\mathbb{R}^d$. By looking at the dependence of the MSD on time, one can distinguish different modes of motion and obtain estimates for the corresponding parameters. Some examples are:

$$\begin{array}{ll} MSD(t) = 2dDt & \text{pure diffusion} \\ MSD(t) = 2dDt^{\alpha} & \text{anomalous diffusion} \\ MSD(t) = 2dDt + (\|V\|t)^2 & \text{directed diffusion} \end{array} \tag{4.4}$$

where $D$ and $V$ are, respectively, the diffusion coefficient and the velocity vector, and $\alpha \neq 1$ is a positive real number. The form of the MSD in Equation (4.4) for pure and directed diffusion is an immediate consequence of Equation (2.2) and Equation (2.3). In Figure 4.4, we plot the behaviour of the MSD for different modes of motion of the particles.

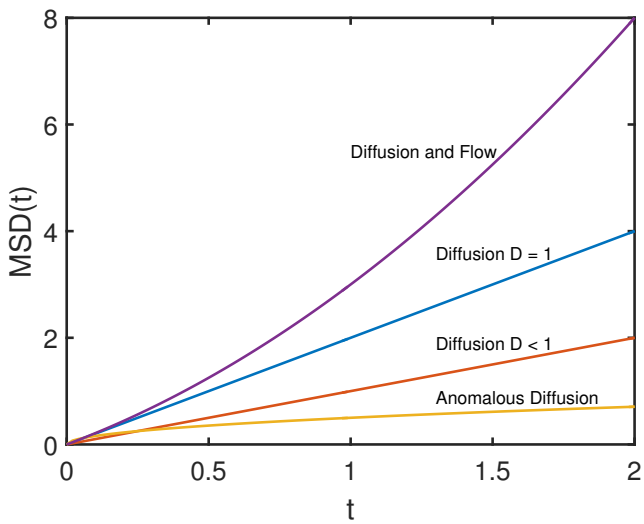


Figure 4.4: The mean square displacement as a function of time for simultaneous diffusion and flow (directed diffusion), pure diffusion with $D = 1$ and $D < 1$, and anomalous diffusion with $D = 0.5$ and $\alpha = 0.5$.

### 4.2.1 Single Particle Raster Image Analysis

In typical SPT experiments, particles move negligibly within an image and appreciably between consecutive images. Thus, the motion is estimated from the position of a particle in consecutive images. In Single Particle Raster Image Analysis (SPRIA), raster images are analysed where the scanning speed is such that the time between adjacent pixels in the $x$-direction is small (the pixel dwell time is in the order of a microsecond) while the time between adjacent pixels in the $y$-direction is large (the line dwell time is in the order of a millisecond). Hence, particles will move between consecutive lines in an image. More details are provided in the two appended papers, where the SPRIA method is introduced, discussed and validated on both simulated and experimental data. In this introduction we only recall briefly the main steps of SPRIA.

A particle is defined by an axis-parallel rectangle through a double threshold method. The first threshold is used to discriminate whether a local maximum of photon counts is an actual particle as opposed to noise, while the second threshold is adopted to delineate the boundary of the rectangle. In Figure 4.5, an identified particle is depicted, where pixels are colored based on their intensity in the image. Moreover, it can immediately be seen that two things in SPRIA are different from a typical SPT experiment: first, particles do not look round anymore as they are allowed to move while we are scanning them, producing a pattern of bright shifted line segments; second, linking the successive positions of the particles to form tracks is more straightforward as the bright lines forming a particle, which corresponds to the different time points of the trajectory, tend to be connected, see Figure 4.5. Once a particle has been extracted as described above, its position in each line, i.e. in each time step of the trajectory, is estimated either by a maximum likelihood method based on the assumption of independently Poisson distributed photon counts in each pixel (Paper I) or by a centroid method (Paper II). In Figure 4.5, the trajectories estimated by both methods together with the true one are shown, indicating that SPRIA works well. Then, an estimate of the diffusion coefficient of the particle is obtained by using Equation (4.4) for pure diffusion when $t$ is set to be the time $\tau_l$ between two consecutive lines. Finally, an overall estimate of one, or more in the case of particle mixtures, diffusion coefficients can be retrieved from the distribution of the diffusion coefficients of the single particles.

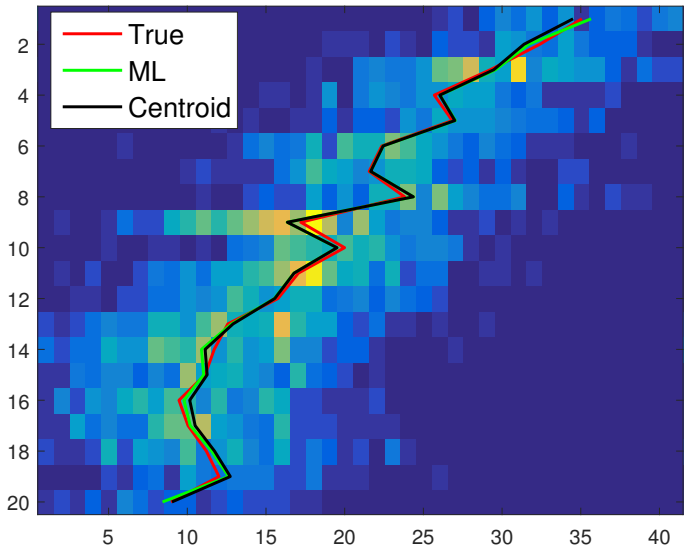


Figure 4.5: A simulated raster scan image of a 50nm particle showing the true trajectory (red), the corresponding estimated trajectory computed using the maximum likelihood method (green) introduced in Paper I, and the centroid based method (black) presented in Paper II.

# 5 Model selection

In many applications, the goal of the statistical analysis is to make inference about the parameters of a model, and evaluate the goodness of fit of different models to choose the best one to explain the observed data. Two widely used applications are regression and mixture models. The former deals with one of the most fundamental problems in science, that is to explain dependencies between variables. The latter is often used to describe a population, the behaviour of which can be characterized by the behaviour of its subgroups. For example, in regression problems we would like to decide how many and which predictors we should include in the model, while in mixture models we would like to determine how many subgroups are present in the population. Standard criteria used in model selection balances a measure of goodness of the fit or predictive power with some form of penalization for the complexity of the model. In general, model selection is an open problem, and various solutions have been proposed for particular models, data types, and sample sizes. In the following discussion, we restrict ourselves to the case of mixture modelling.

Consider a parametric family of distributions with corresponding density function $g(\cdot, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters for the distribution. For example, if $g$ is the density function of a normal distribution, then $\boldsymbol{\theta} = (\mu, \Sigma)$ is the vector containing the mean and covariance matrix of the random variable. Assume we have a random sample $X_1, \ldots, X_n$ of size $n$ with a density function $f$ given by

$$f(x, \boldsymbol{\theta}) = \sum_{j=1}^{k_{\text{true}}} p_j g(x, \boldsymbol{\theta}_j), \quad \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_{\text{true}}}, p_1, \ldots, p_{k_{\text{true}}}) \tag{5.1}$$

where $p_j$ are the mixing coefficients and satisfy the conditions $0 \leq p_j \leq 1$, and $\sum_{j=1}^{k_{\text{true}}} p_j = 1$. The parametric family $g(\cdot, \boldsymbol{\theta})$ is used to model the behaviour of the subpopulations, and by allowing different parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_{\text{true}}}$ we can capture the behaviour of the $k_{\text{true}}$ subgroups. The mixing coefficient $p_j$ indicates how likely it is that an observation from the population belongs to the subgroup $j$. The model in Equation (5.1) can be rewritten in an equivalent way by introducing for each observation $X_1, \ldots, X_n$ the corresponding random variable $Z_1, \ldots, Z_n$ indicating the group membership, i.e. distributed as $Z = j$ if $X$ belongs to group $j$. We can then interpret the mixing coefficients as $p_j = P(Z = j)$ and $g(\cdot, \boldsymbol{\theta}_j)$ as the conditional distribution of $X$ given that it belongs

to the $j$-th group. The variables $Z$ are usually called latent variables as most often they are unobservable. The goal of the statistical analysis in this context is twofold: first, estimate the parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_{\text{true}}}$ of the distributions of the different groups and their proportions $p_1, \ldots, p_{k_{\text{true}}}$; second, to identify how many different groups are present in the population, that is, to estimate $k_{\text{true}}$. The parameter estimation problem is typically solved using maximum likelihood method. However, since the latent variables are not observed, the maximization of the likelihood is difficult. Thus, we typically fit the mixture model by using the Expectation Maximization (EM) algorithm (Dempster *et al.*, 1977). The idea behind the EM algorithm is that the estimation problem would be simplified if the group memberships were known. In fact, if that was the case the complete log-likelihood for a mixture model with $K$ components would become

$$\log(f(\boldsymbol{\theta}|X_1, \ldots, X_n, Z_1, \ldots, Z_n)) = \sum_{j=1}^{n} \sum_{k=1}^{K} \mathbb{1}(Z_j = k) \log(p_j g(X_j, \boldsymbol{\theta}_k)).$$

Briefly, the EM algorithm proceeds as follows: choose an initial guess $\boldsymbol{\theta}^0 = (\boldsymbol{\theta}_1^0, \ldots, \boldsymbol{\theta}_{k_{\text{true}}}^0, p_1^0, \ldots, p_{k_{\text{true}}}^0)$ for the parameters. Iterate the following two steps for $m = 1, 2, \ldots$ until some chosen convergence criterion is met:

**Expectation step:** Compute the conditional expectation of the log-likelihood with respect to the latent variables

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m-1)}) &= \mathbb{E}[\log(f(\boldsymbol{\theta}|X_1, \ldots, X_n, Z_1, \ldots, Z_n))|X_1, \ldots, X_n, \boldsymbol{\theta}^{(m-1)}] \\ &= \sum_{j=1}^{n} \sum_{k=1}^{K} P(Z_j = k|X_1, \ldots, X_n, \boldsymbol{\theta}^{(m-1)}) \log(p_j g(X_j, \boldsymbol{\theta}_k)), \end{aligned}$$

where

$$P(Z_j = k|X_1, \ldots, X_n, \boldsymbol{\theta}) = \frac{p_k g(X_j, \boldsymbol{\theta}_k)}{\sum_{r=1}^{K} p_r g(X_j, \boldsymbol{\theta}_r)}$$

is the posterior probability that observation $j$ belongs to the group $k$.

**Maximization step:** Update the parameter vector $\boldsymbol{\theta}^m$ by

$$\boldsymbol{\theta}^m = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m-1)}).$$

One of the reasons why the EM algorithm is very popular is that, under some mild regularity conditions, it is guaranteed to converge to a local maximum. Typically, we run the algorithm multiple times with different initial guesses for the parameters to ensure convergence to the global maximum. As the two steps

of the EM algorithm often do not admit an analytical solution, some generalizations of this algorithm have been proposed. For example, the expectation step can be solved using Monte Carlo methods, while the maximization step can be performed numerically with the steepest descent method. Moreover, the EM algorithm is not limited to the maximum likelihood method, and is often used in Bayesian approaches. The probabilities $P(Z_j = k|X_1, \ldots, X_n, \boldsymbol{\theta})$ obtained from the EM algorithm can be used to classify the observations. In the context of this thesis, classification is not the main goal of the investigations, as in the application of diffusing particles we are not concerned with the classification of the individual particles, but rather to understand how many different sizes of particles are present in the sample. Suppose now that mixture models with different numbers of components $K = 1, 2, ...k_{\max}$ have been fitted with the EM algorithm. One is then faced with the problem of choosing the correct number of components as supported by the empirical evidence. While the log-likelihood measures the goodness of fit of these models, it cannot be directly used to estimate the order of the mixture model $k_{\text{true}}$. In fact, the log-likelihood is an increasing function of the number of components as more complex models will always provide a better fit to the data. There are several ways in which one can deal with the overfitting of the likelihood. A very popular solution is to add a penalization term to the log-likelihood. In this family, we find the well-studied Akaike information criterion (AIC) Akaike (1970) and Bayesian information criterion (BIC) Schwarz (1978). In AIC, we maximize the function $\log L - n_{\text{par}}$, where $\log L$ is the log-likelihood and $n_{\text{par}}$ is the number of model parameters. In BIC, we maximize $\log L - 0.5 n_{\text{par}} \log n$ with $n$ equal to the number of observations. Another approach to this problem is to use cross-validation. In cross-validation the observations are divided into a training set and a validation set. The models are then fitted on the training set and their performance assessed in terms of predictive power on the validation set. The type of split of the observations in the training and validation sets defines the type of cross-validation. For example, in leave-one-out cross-validation, in turn each observation is used as the validation set and the remaining $n - 1$ are used as the training set. In V-fold cross-validation the dataset is divided into $V$ disjoint sets of (approximately) the same size $n/V$ and each of these sets is used in turn as the validation set. In this thesis, we implement Monte Carlo cross-validation (MCCV). In MCCV, the observations are repeatedly and randomly divided into a training set of size $n_t$ by drawing observations from the data without replacement, and the remaining observations form a validation set. The choice of the number of components, both when using information criteria and cross-validation, is made by maximizing the measure of the goodness of fit considered.

# 6 Summary of papers

In this section, we introduce the methods used and summarise the results presented in the four appended papers. Regarding the simulated data analysed in this thesis, diffusion was reproduced by simulating discrete time Brownian motion of spheres in a box with periodic boundary conditions. Different settings of the scan rate, pixel size, pixel dwell time and line time were considered. A calibration step was performed on immobilized 175 nm beads in gelatin to obtain the lateral and axial waists of the point spread function for our experimental setup.

## 6.1 Paper I: Single particle raster image analysis of diffusion

The introduction of raster image correlation spectroscopy has lead to a shift in the spatiotemporal analysis of dynamics in complex heterogeneous systems. By exploiting the time structure within single raster images, it is possible to increase the time resolution and resolve dynamics at shorter timescales by means of the quick pixel-to-pixel sampling. In this article, we introduced Single Particle Raster Image Analysis (SPRIA), a single particle method to study raster images. The motivation of this study was to develop a method that could locally map mass transport properties. Previously, RICS has been applied to study heterogeneity Schuster *et al.* (2016), however, as it gains strength from the averaging of many molecules, its spatial resolution is limited by the minimum size of the region of interest. In SPRIA, single particles are extracted using a double threshold method, where one thresholding is used to define which local maxima are particles and another to separate the particles from the background. The maximum likelihood method is then employed to reconstruct the tracks of the molecules based on the assumption of pixelwise independent Poisson distributed photon counts. Two main problems were encountered when developing SPRIA: first, the symmetry of the likelihood with respect to the $y$ and $z$ coordinates made us restrict ourselves to estimate the diffusion coefficient only from the motion along the $x$-axis. Second, the raster scanning introduces a bias on the observed diffusion coefficient which is more significant the slower the scan rate is. This is due to the inherent preferential sampling of small compared to large line-to-line displacements. We suggested a simulation based method to correct for this bias. Both on simulated and experimental data, SPRIA has shown to provide accurate estimates as compared to RICS. In the simulation study, we

demonstrated that using the bias correction leads to better estimates. Finally, we introduced a bootstrap method to estimate standard errors in RICS, where images are resampled from the original stack of images to create new datasets. The motivation behind the introduction of the bootstrapped standard error comes from the observation that in some cases, the traditional way of estimating standard errors for RICS by means of the residuals gives unrealistically small values. The explanation for such small estimates could be that the residuals are highly correlated.

## 6.2 Paper II: Raster image analysis of diffusion for particle mixtures

In this study we extended the work done in Paper I to mixtures of particles. In SPRIA, the motion of each single particle is estimated, and we gain information about the distribution of the mean square displacement and functions that depend on it. As the mathematical model used for pure diffusion corresponds to a particle performing a Brownian motion, the theoretical distribution of the estimated diffusion coefficients can be computed, and involves a gamma distribution with parameters depending on the true diffusion coefficient and the length of the observed trajectory. We set up a maximum likelihood method to detect mixture models and estimate the diffusion coefficients of the different populations. In the validation study, SPRIA has been shown to give good estimates, but some caution must be taken when selecting the number of components in the mixture. When using criteria based on likelihood improvement, the maximum suggested number of components in the mixture is always selected, indicating that the likelihood is too sensitive to the variability in the distribution of the diffusion coefficient. Thus, a complementary condition, where we rejected components for which the estimated proportion fell below a threshold, was necessary. We also found that when applying RICS to such complex systems of mixtures of particles, a rather large difference between the components in the mixture was needed to allow for identifiability. Thus, we investigated the use of RICS for mixtures by looking in more detail at the correlation function for different models involving diffusion to quantify how large the difference between the diffusion coefficients of the components need to be to be able to see a difference in the correlation functions.

## 6.3 Paper III: Identification of mixture models with an application to diffusing particles

The results and the application in Paper II motivated us to investigate model selection criteria based on cross-validation. In cross-validation, the observations

are divided into a training set, used to estimate the parameters of the models, and a validation set, used to compare the different models according to some goodness of fit measure. We use in our analysis Monte Carlo cross-validation, where we form the training set by drawing without replacement $n_t$ data points from the $n$ observations available. Here, model selection in the context of mixture models is investigated by means of simulations and experiments. Since the likelihood provides a direct way to measure the performance of different models but, as presented in Paper II, tends to choose more complex models than the true one, we consider different penalizations. The results from cross-validation are compared to results based on the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) which both penalize the likelihood of a model by its complexity, as measured by the number of parameters in the model, and for BIC the penalization gets stronger as more observations become available. In addition, we consider the cross-validated likelihood, cross-validated AIC, and cross-validated BIC with different proportions of observations in the training and validation sets. Moreover, we include cross-validation with the one standard error rule that penalizes even further the complexity by taking into account the variability of the cross-validated measure. We found that, for mixtures of normal distributions, the cross-validated log-likelihood with a carefully chosen proportion of observations in the training set outperforms both AIC and BIC. The cross-validated log-likelihood seems to reliably select the correct number of components in mixtures with one, two, and three components and in the case of both even or uneven proportions. We present a finite sample bound for the error between the cross-validated conditional risk and the oracle benchmark conditional risk. Surprisingly, for mixtures of gamma distributions, AIC overall surpasses all the other criteria considered. For both normal and gamma mixtures, the sample size plays a major role in the performance of the different criteria used to perform model selection. We reviewed the results for the SPRIA experiments presented in Paper II. We tested AIC, BIC and different cross-validated measures and none of them turned out to be a proper model selection rule. Thus, we investigated further how some constraints on the parameters of the mixture models, in particular the mixing coefficients, could help in model selection. All criteria led to models with spurious components with relatively low proportions, and we decided to impose a lower bound for the proportions in the model and assessed the performance of different thresholds. In conclusion, selecting the best model according to either AIC or BIC such that all its components have a minimum proportion of 15% worked relatively well for the experiments.

## 6.4 Paper IV: Statistics of Raster Image Correlation Spectroscopy

Diffusing particles observed with a confocal laser scanning microscope give rise to a doubly stochastic Poisson point process. In particular, the photon detected by the microscope in one pixel follows a Poisson distribution with a parameter that depends on the particle positions in space, which are modelled as a Poisson point process. The technique Raster Image Correlation Spectroscopy is based on the statistics of the photon detection process and has been increasingly applied to study molecular transport in cells and solutions. Our approach was inspired by the work of Qian (1990) and Koppel (1974), where they looked at the asymptotic behaviour of the signal to noise ratio of fluorescence correlation spectroscopy under different assumptions. In our study, we have tried to bridge the theoretical analysis of RICS as a statistical method with its application in experiments. We show that the moments of the photon detection process can be computed in terms of physically relevant parameters such as the diffusion coefficient of the particles, their brightness and others. As a direct consequence, the statistical accuracy of the above mentioned technique can be evaluated. We propose the method called Raster Image Correlation Spectroscopy Performance Evaluation (RICSPE) to compare different experimental setups. RICSPE examines the distribution of the estimated diffusion coefficient given a set of experimental parameters. Thus, we can relate the different experimental parameters that affect the photon detection process to the accuracy of RICS, allowing us to optimally design an experiment. It has been claimed for a long time that the results from the RICS analysis would depend on the scan speed, and that for a successful experiment the pixel dwell time must be appropriate for the diffusion coefficient being measured. For the first time, we uncovered the dependency of the estimated diffusion coefficient on the scan speed, and quantified this effect. The important parameter for slow diffusing particles ($D = 0.1 - 10\mu m^2 s^{-1}$) is the line time, while it is the pixel dwell time for fast diffusing particles ($D = 100 - 400\mu m^2 s^{-1}$). The brightness, which can be controlled for example by setting the laser power, should be in the range of $10^5 - 10^7$ counts per particle per second for solution experiments, and $10^4$ for cell measurements. Higher brightness would lead to photobleaching and possibly saturation effects. We found that the image size should be at least $200 \times 200$ pixels, and the pixel size should be at least four times smaller than the radius of the point spread function. We tested our findings against simulations and experiments and summarized our conclusions in ready-to-use guidelines. To further promote RICSPE and RICS, a graphical user interface for the algorithms developed here is made available through a repository.

# 7 Conclusions and future work

The work behind this thesis had two goals: to develop new statistical methods to study mass transport properties, such as diffusion, both in homogeneous and heterogeneous systems and to further improve existing microscopy methods. To achieve this, we have investigated simulations and experiments performed with a confocal laser scanning microscope on monodisperse and polydisperse systems of diffusing particles. More specifically, in Paper I we have introduced the method called SPRIA, which is an adaptation of single particle tracking to the case of raster images, and compared it to the widely used RICS technique. Moreover, we showed with a proof of principle simulation that SPRIA could produce a mobility map by estimating locally mass transport properties in a heterogeneous sample with two regions with different viscosities. This result, which is not achievable with the standard RICS method, together with the results in Paper II motivated the introduction of SPRIA. In fact, in Paper II we concluded the diffusion coefficients of different probes need to differ by a factor 8 to reliably estimate the parameters of the model with RICS, in case no a priori information is available about the number of components in the mixture. On the other hand, SPRIA performs satisfactorily even when we have no prior knowledge of the sample. Furthermore, we tested model selection by means of likelihood ratio test and concluded that, in particular for experimental data, this criterion could not alone be used to determine the number of components in the mixture as spurious components were added. Thus, we proposed to constrain the mixing coefficients of the mixture model with a threshold of 15%. We studied further this problem in Paper III. The conclusions therein showed that for experiments of polydisperse systems of particles the 15% rule proved to be the most satisfactory when compared to standard criteria as AIC, BIC, and Monte Carlo cross-validation. The model selection criteria considered can be utilized in a wide range of applications, as we showed that cross-validation outperforms AIC and BIC for normal mixtures. Normal mixture models have been used both in density estimation problems and cluster analysis, two of the most studied statistical problems in the literature. The main contribution of this thesis to the improvement of existing microscopy methods deals with the statistical analysis of RICS. We described the effect of both experimental and molecular parameters in detail, such as the scan speed, concentration, molecular brightness, and pixel size on the performance of RICS. The theory behind the RICSPE method that was developed is novel and is of general relevance to the confocal microscopy literature. The recommendations and guidelines provided

to design a RICS experiment are unique and they will help improve the work of the experimentalists. The statistical methods introduced are not limited to RICS but can be adapted for any ICS method.

A natural continuation of the current work would be the extension of the finite sample theoretical result in Paper III to a mixture model with a parametric family of distributions different from normal densities, as it would be very relevant for the statistical community. As the experiments in Paper II suggest, it would be interesting to investigate how model mispecifications, such as having noisy observations or an approximate likelihood, affect clustering and model selection. As for SPRIA, it would be interesting to apply this new method and compare it with the ICS techniques for heterogeneous samples with spatially varying mass transport properties and possibly with interacting particles. Concerning the application on confocal microscopy, it would be advantageous to consider more complex and realistic models by including both non-stationary effects caused by long measurement times, such as photobleaching, and preprocessing of the images, for example the use of a moving average to filter immobile artifacts.

# References


Akaike, H. (1970) Statistical predictor identification. *Ann. Inst. Statist. Math.*, **22**, 203–217.

Bouchard, J. & Georges, A. (1990) Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications. *Physics Reports*, **195**, 127–293.

Brown, C. M., Dalal, R. B., Herbert, B., Digman, M. A., Horwitz, A. R. & Gratton, E. (2008) Raster image correlation spectroscopy (RICS) for measuring fast protein dynamics and concentrations with a commercial laser scanning confocal microscope. *Journal of Microscopy*, **229**, 78–91.

Comeau, J. W., Costantino, S. & Wiseman, P. W. (2006) A guide to accurate fluorescence microscopy colocalization measurements. *Biophysical Journal*, **89**, 1251–1260.

Cox, D. R. (1955) Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, **17**, 129–164.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B*, **39**, 1–38.

Diggle, P. (2013) *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition*. CRC Press, London.

Digman, M. A., Brown, C. M., Sengupta, P., Wiseman, P. W., Horwitz, A. R. & Gratton, E. (2005) Measuring fast dynamics in solutions and cells with a laser scanning microscope. *Biophysical Journal*, **89**, 1317–1327.

Einstein, A. (1905) Über die von molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, **322**, 549–560.

Gielen, E., Smisdom, N., Vandeven, M., De Clercq, B., Gratton, E., Digman, M., Rigo, J.-M., Hofkens, J., Engelborghs, Y. & Ameloot, M. (2009) Measuring diffusion of lipid-like probes in artificial and natural membranes by raster image correlation spectroscopy (RICS): use of a commercial laser-scanning microscope with analog detection. *Langmuir*, **25**, 5209–5218.

Hebert, B., Costantino, S. & Wiseman., P. W. (2005) Spatiotemporal image correlation spectroscopy (STICS) theory, verification, and application to protein velocity mapping in living CHO cells. *Biophysical Journal*, **88**, 3601–3614.

Kolin, D. L., Ronis, D. & Wiseman, P. W. (2006) K-space image correlation spectroscopy: a method for accurate transport measurements independent of fluorophore photophysics. *Biophysical Journal*, **91**, 3061–3075.

Koppel, D. E. (1974) Statistical accuracy in fluorescence correlation spectroscopy. *Phys. Rev. A*, **10**, 1938–1945.

Kulkarni, R., Wu, D., Davis, M. E. & Fraser, S. E. (2005) Quantitating intracellular transport of polyplexes by spatio-temporal image correlation spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 7523–7528.

Lorén, N., Hagman, J., Jonasson, J. K., Deschout, H., Bernin, D., Cella-Zanacchi, F., Diaspro, A., McNally, J. G., Ameloot, M., Smisdom, N., Nydén, M., Hermansson, A.-M., Rudemo, M. & Braeckmans, K. (2015) Fluorescence recovery after photobleaching in material and life sciences: putting theory into practice. *Quarterly Reviews of Biophysics*, **48**, 323–387.

Pawley, J. B. (2006) *Handbook of Biological Confocal Microscopy*. Springer Science, New York.

Perrin, J. (1909) Mouvement brownien et réalité moléculaire. *Ann. Chem. Phys.*, **18**, 5–114.

Qian, H. (1990) On the statistics of fluorescence correlation spectroscopy. *Biophysical Chemistry*, **38**, 49 – 57.

Saleh, B. (1978) *Photoelectron Statistics, with Applications to Spectroscopy and Optical Communication*. Springer-Verlag Berlin ; New York.

Schuster, E., Sott, K., , Ström, A., Altskär, A., Smisdom, N., Lorén, N. & Hermansson, A.-M. (2016) Interplay between flow and diffusion in capillary alginate hydrogels. *Soft Matter*, **12**, 3897–3907.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Smoluchowski, M. (1906) Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen. *Annalen der Physik*, **326**, 756–780.